Exercise 1: Lasso Regularization

Consider the regression learning setting, i.e., $\mathcal{Y} = \mathbb{R}$, and feature space $\mathcal{X} = \mathbb{R}^p$. Let the hypothesis space be the linear models:

$$\mathcal{H} = \{f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{X} \mid \boldsymbol{\theta} \in \mathbb{R}^p\}.$$

Suppose your loss function of interest is the L2 loss $L\left(y, f(\mathbf{x})\right) = \frac{1}{2}\big(y - f(\mathbf{x})\big)^2$. Consider the L_1-regularized empirical risk of a model $f(\mathbf{x} \mid \boldsymbol{\theta})$ (i.e., Lasso regression):

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1 = \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{X}^{(i)}\right)^2 + \lambda \sum_{i=1}^{p} |\theta_i|.$$

Assume that $\mathbf{X}^T\mathbf{X} = \boldsymbol{I}$, which holds if $\mathbf{X}$ has orthonormal columns. Show that the minimizer $\hat{\theta}_{\text{Lasso}} = (\hat{\theta}_{\text{Lasso},1}, \ldots, \hat{\theta}_{\text{Lasso},p})^\top$ is given by

$$\hat{\theta}_{\text{Lasso},i} = sgn(\hat{\theta}_i) \max\{|\hat{\theta}_i| - \lambda, 0\}, \quad i = 1, \ldots, p,$$

where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^\top = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the minimizer of the unregularized empirical risk (w.r.t. the L2 loss). For this purpose, use the following steps:

(a) Derive that

$$\arg\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{p} -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda|\theta_i|.$$

(b) Note that the minimization problem on the right-hand side of (a) can be written as $\sum_{i=1}^{p} g_i(\theta_i)$, where

$$g_i(\theta_i) = -\hat{\theta}_i \theta_i + \frac{\theta_i^2}{2} + \lambda|\theta_i|.$$

What is the advantage of this representation if we seek to find the $\boldsymbol{\theta}$ with entries $\theta_1, \ldots, \theta_p$ minimizing $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$?

(c) Consider first the case that $\hat{\theta}_i > 0$ and infer that for the minimizer θ_i^* of g_i it must hold that $\theta_i^* \geq 0$.
Hint: Compare $g_i(\theta_i)$ and $g_i(-\theta_i)$ for $\theta_i \geq 0$.

(d) Derive that $\theta_i^* = \max\{\hat{\theta}_i - \lambda, 0\}$, by using (c) (and also still considering the case $\hat{\theta}_i > 0$.)

(e) Consider the complementary case of (c) and (d), i.e., $\hat{\theta}_i \leq 0$, and infer that for the minimizer θ_i^* of g_i it must hold that $\theta_i^* \leq 0$.

(f) Derive that $\theta_i^* = \min\{\hat{\theta}_i + \lambda, 0\}$, by using (e) (and also still considering the case $\hat{\theta}_i \leq 0$.)

(g) Make sure that both minimizers in the two case can indeed be written as $sgn(\hat{\theta}_i)\max\{|\hat{\theta}_i| - \lambda, 0\}$.